DOCUMENT RESUME

ED 416 214 TM 028 066

AUTHOR Thompson, Bruce
TITLE Why "Encouraging" Effect Size Reporting Isn't Working: The Etiology of Researcher Resistance to Changing Practices.
PUB DATE 1998-01-22
NOTE 18p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Houston, TX, January 1998).
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Attitudes; Change; Denial (Psychology); *Educational Research; *Effect Size; *Etiology; *Research Methodology; *Researchers; *Statistical Significance

ABSTRACT

Given decades of lucid, blunt admonitions that statistical significance tests are often misused, and that the tests are somewhat limited in utility, what is needed is less repeated bashing of statistical tests, and some honest reflection regarding the etiology of researchers' denial and psychological resistance (sometimes unconscious) to improved practice. Three etiologies are briefly explored here: (1) atavism; (2) "is/ought" logic fallacies; and (3) confusion/desperation. Understanding the etiology of psychological resistance may ultimately lead to improved interventions to assist in overcoming researcher resistance to reporting effect sizes and using non-nil nulls and other analytic improvements. (Contains 45 references.) (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

apsstsig.wp1 1/20/98

ED 416 214

WHY "ENCOURAGING" EFFECT SIZE REPORTING ISN'T WORKING: THE ETIOLOGY OF RESEARCHER RESISTANCE TO CHANGING PRACTICES

Bruce Thompson

Texas A&M University 77843-4225
and
Baylor College of Medicine

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Bruce Thompson

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM028066

Paper presented at the annual meeting of the Southwest Educational Research Association, Houston, January 22, 1998. The impetus for writing the present paper originated from the thoughtful comment of Mark Applebaum on a draft of my 1997 APA Invited Address. The author may be contacted through Internet URL: http://acs.tamu.edu/-bbt6147/.

BEST COPY AVAILABLE

ABSTRACT

Given decades of lucid, blunt admonitions that statistical significance tests are often misused, and that the tests are somewhat limited in utility, what is needed is less repeated bashing of statistical tests, and some honest reflection regarding the etiology of researchers' denial and psychological resistance (sometimes unconscious) to improved practice. Three etiologies are briefly explored here: (a) atavism, (b) "is/ought" logic fallacies, and (c) confusion/desperation. Understanding the etiology of psychological resistance may ultimately lead to improved interventions to assist in overcoming researcher resistance to reporting effect sizes and using non-nil nulls and other analytic improvements.

BEST COPY AVAILABLE

A number of clarion calls have been published urging researchers to abandon, or to at least supplement, the use of statistical significance tests (e.g., Hunter, 1997; Kirk, 1996; Schmidt, 1996; and Thompson, 1996, 1997a). Indeed, articles in this genre are so common that "it is more difficult to find specific arguments for significance tests than it is to find arguments decrying their use" (Henkel, 1976, p. 87; but see selected chapters in Harlow, Mulaik & Steiger, 1997).

For example, the lead section of the January, 1997 issue of Psychological Science was devoted to this controversy. A seemingly periodic series of articles on the extraordinary limits of statistical significance tests has been published in the American Psychologist (cf. Cohen, 1990, 1994; Kupfersmid, 1988; Rosenthal, 1991; Rosnow & Rosenthal, 1989). The entire 1993 Volume 61, Number 4 issue of the Journal of Experimental Education dealt with these themes. The April, 1998 issue of Educational and Psychological Measurement featured two lengthy reviews (cf. Thompson, in press) of a major text (Harlow, Mulaik & Steiger, 1997) on the controversy. And the APA Task Force on Statistical Inference (Shea, 1996) has now been working for more than a year on related recommendations for improving practices.

These numerous articles have advanced various themes, but three themes stand out. First, statistical significance tests do not evaluate result importance or value (cf. Thompson, 1993). Second, statistical significance tests do not evaluate result replicability (cf. Cohen, 1994). Third, statistical significance

tests do not measure result magnitude or effect (cf. Thompson, 1997b).

Unhappily, journal editor Loftus (1994), like others, has lamented that repeated publications of

> these concerns never seem to attract much attention (much less impel action). They are carefully crafted and put forth for consideration, only to just kind of dissolve away in the vast acid bath of our existing methodological orthodoxy. (p. 1)

Another editor commented: "p values are like mosquitos" that apparently "have an evolutionary niche somewhere and [unfortunately] no amount of scratching, swatting or spraying will dislodge them" (Campbell, 1982, p. 698).

Similar comments have been made by non-editors. For example, Falk and Greenbaum (1995) noted that "A massive educational effort is required to... extinguish the mindless use of a procedure that dies hard" (p. 94). Harris (1991) observed, "it is surprising that the dragon will not stay dead" (p. 375). And recently Rozeboom (1997) argued:

> Null-hypothesis significance testing is surely the most bone-headed misguided procedure ever institutionalized in the rote training of science students... [I]t is a sociology-of-science wonderment that this statistical practice has remained so unresponsive to criticism... (p. 335)

Particularly disturbing has been resistance to following the

admonitions of the new APA style manual, which noted that:

> Neither of the two types of probability values [statistical significance tests] reflects the importance or magnitude of an effect because both depend on sample size... You are [therefore] *encouraged* to provide effect-size information. (APA, 1994, p. 18, emphasis added)

However, *empirical* studies of articles published since 1994 in psychology, counseling, special education, and general education suggest that merely "*encouraging*" effect size reporting (APA, 1994) has <u>not</u> appreciably affected actual reporting practices (e.g., Kirk, 1996; Snyder & Thompson, in press; Thompson & Snyder, 1997, in press; Vacha-Haase & Nilsson, in press).

The present survey explores the etiology of this resistance to change. Schmidt and Hunter (1997), for example, have suggested that it may be necessary to invoke psychological principles to explain researcher resistance to change. They noted that, "logic-based arguments seem to have had only a limited impact... [perhaps due to] the virtual brainwashing in significance testing that all of us have undergone" (Schmidt & Hunter, 1997, pp. 38-39). Schmidt and Hunter (1997) also spoke of a "psychology of addiction to significance testing" and acknowledged that "changing the beliefs and practices of a lifetime... naturally... provokes resistance" (Schmidt & Hunter, 1997, p. 49).

Three etiologies of psychological resistance, which may be either conscious or unconscious, are briefly explored here: (a)

atavism, (b) "is/ought" logic fallacies, and (c) confusion/desperation. Understanding the etiology of this psychological resistance may ultimately lead to improved interventions to assist in overcoming researcher resistance to reporting effect sizes (Kirk, 1996; Snyder & Lawson, 1993) and using non-nil nulls (see Cohen (1994), Meehl (1997), and Thompson (in press)) and other improvements. The ideas presented here are not empirically validated, but do represent musings that delineate a constellation of possible explanations for observed resistance/denial phenomena.

Etiology #1: *Atavism*

Existential psychologists posit that an atavistic desire to escape freedom and responsibility underlies much human behavior, including presumably the behavior of the humans called researchers. Some researchers inappropriately feel that they can equate an improbable result with an inherently important result (see Shaver, 1985), and thereby finesse the responsibility for and necessity of declaring and exposing to criticism the personal or societal values that inherently must be the basis for any decree that research results are valuable (see Thompson, 1993).

In this vein some researchers "escape from freedom" (see Fromm's book with that title) and responsibility by asserting that they have no control over the normative scholarly practices of their field, and therefore little or no responsibility for their own behaviors. For example, in his recent defense of statistical significance tests, Hagen (1997) argued that, "It is unlikely that

we will ever be able to divorce ourselves from that [statistical test] logic even if someday we decide that we want to" (p. 22).

Researchers acting under the purview of this model say things such as, "I would like to report an effect size, since statistical significance tests do not evaluate result importance (cf. Thompson, 1993), but I am afraid to deviate from the normative behavior, and most researchers today still do not report effect sizes (cf. Kirk, 1996)." Or researchers may say, "I would like to present some evidence that my results will replicate, since statistical significance tests do not (do not, do not...) evaluate result replicability (cf. Cohen, 1994; Thompson, 1996), but I am afraid to report such results, since most people do not do so, and I am afraid my manuscript will be rejected if I do anything unusual, albeit correct."

Mahoney (1976) provides an example of these dynamics:

> Even though I am very critical of statistical inference... I shall probably continue to pay homage to "tests of significance" in the papers I submit to psychological journals. My rationale for this admitted hypocrisy is straightforward: until the rules of the science game are changed, one must abide by at least some of the old rules or drop out of the game. (p. xiii)

Etiology #2: *"Should/Would" or "Is/Ought" Logic Fallacy*

In his defense of statistical significance tests, Frick (1996) cited an anonymous reviewer of his manuscript who argued that, "A

BEST COPY AVAILABLE

way of thinking that has survived decades of ferocious attacks is likely to have some value" (p. 379). It is ironic that this model invokes exactly the same "should/would" or "is/ought" logic error (Hudson, 1969) made by some critics of statistical significance tests. As Strike (1979) explained, "To deduce a proposition with an 'ought' in it from premises containing only 'is' assertions is to get something in the conclusion not contained in the premises, something impossible in a valid deductive argument" (p. 13).

Some scholars (cf. Carver, 1978) have called for the banning of statistical significance tests, illogically arguing that because these tests "are" so commonly misused, therefore these tests "should" be abandoned. A logically valid related argument would *instead* assert, because statistical tests "should" not be incorrectly interpreted, researchers "should" change their behavior and correctly interpret these tests. On the other hand, advocates of statistical significance tests who argue that, because these tests "are" commonly used, the statistical significance tests *ipso facto* must or "should" be valuable, are themselves merely presenting a "should/would" fallacy in alternative clothing.

Etiology #3: *Confusion/Desperation*

Some combination of confusion and desperation may also explain the resistance of some researchers to changing their analytic practices. Regarding confusion, Biskin (in press) for example argued that "tests of significance are intended for making inferences about populations, not samples." This is exactly the misconception that has led to continuing overreliance on

statistical significance tests.

Carver (1978) characterized the meaning of $p_{CALCULATED}$, the most fundamental element of hypothesis testing, as "the most important and least understood principle of statistical significance testing" (p. 384). The logic of null hypothesis testing is so convoluted that it is small wonder so many researchers remain confused.

As Cohen (1994) and Thompson (1996) explained in detail, statistical significance tests presume that the null hypothesis is exactly true in the population, and then compute the probability (p) of the sample results, given those results and sample size. That is, the direction of the inference in inferential statistics is *from* the population *to* the sample, and <u>not</u> from the sample to the population (Thompson, 1997a, 1997b). Of course, this is not what researchers want statistical significance tests to do!

Researchers want statistical significance tests to evaluate the population, because researchers at a primordial level of psyche dread the embarrassment of discovering the psychological analog of cold fusion--such discoveries lead to one very fun conference involving adulation followed by a lifetime of conferences involving being shunned by all one's peers. If inferential tests did evaluate the population, then we could deduce what future researchers drawing samples from the same population would find. That is, if statistical significance yielded inferences about the population, then these tests would provide useful information about result replicability.

Unfortunately, statistical significance tests do not (do not,

do not...) provide useful information about result replicability, and therefore do not really meet researchers' desperate felt need to avoid embarrassment. Schmidt and Hunter (1997) noted that:

> Psychologically, it is easy to understand the desire for a technique that would perform the desirable function of distinguishing in our data sets between relations, differences, and effects that are real and those that are just chance fluctuations... But wanting to believe something is true does not make it true... (p. 42)

Similarly, Cohen (1994) noted that even though the statistical significance test "does not tell us what we want to know, ...we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!" (p. 997).

Summary

Nearly 40 years ago Nunnally (1960) noted, "We should not feel proud when we see the psychologist smile and say 'the correlation is significant beyond the .01 level.' Perhaps this is the most that he [sic] can say, but he has no reason to smile" (p. 649). In that same year, Rozeboom (1960) observed that "the perceptual defenses of psychologists are particularly efficient when dealing with matters of methodology, and so the statistical folkways of a more primitive past continue to dominate the local scene" (p. 417).

Several scholars (cf. Thompson, 1997b, in press) have suggested that psychological resistance to changing practices will not be overcome until editorial policies *require* such changes. For

BEST COPY AVAILABLE

example, Reichardt and Gollob (1997) argued that

> ...[W]e believe a substantial increase in the use of confidence intervals is unlikely to occur unless substantial changes are made in the process by which submitted articles are accepted or rejected for publication. (p. 279)

That is, the APA (1994) "encouragement" of correct practice will simply not be sufficient. Several *empirical* studies of volumes from diverse journals published since 1994 seem to corroborate this view (Kirk, 1996; Snyder & Thompson, in press; Thompson & Snyder, 1997, in press; Vacha-Haase & Nilsson, in press).

Fortunately, independent of the APA style manual requirements, some journal editors have now written journal-specific policies that are considerably more enlightened. For example, the author guidelines of the Journal of Experimental Education indicate that "authors are required to report and interpret magnitude-of-effect measures in conjunction with every *p* value that is reported" (Heldref Foundation, 1997, pp. 95-96, emphasis added). The author guidelines for Educational and Psychological Measurement are equally informed:

> We will go further. Authors reporting statistical significance will be required to both report and interpret effect sizes. However, these effect sizes may be of various forms, including standardized differences, or uncorrected (e.g., r^2, R^2, eta^2) or corrected (e.g., adjusted R^2, omega^2) variance-

> accounted-for statistics. (Thompson, 1994, p. 845, emphasis in original)

Indeed, there has been movement even among APA journal editors. The editorial policies of the Journal of Applied Psychology now indicate that

> If an author decides not to present an effect size estimate along with the outcome of a significance test, I will ask the author to provide specific justification for why effect sizes are not reported. So far, I have not heard a good argument against presenting effect sizes. Therefore, unless there is a real impediment to doing so, you should routinely include effect size information in the papers you submit. (Murphy, 1997, p. 4)

We can all hope (and work) for the final dawning of some future day when researchers begin to focus on evidence that their results involve (a) noteworthy effects that (b) replicate under stated conditions. To facilitate this day's arrival, we must understand researcher resistance to changing practices. Fortunately, some change is beginning to occur.

References

American Psychological Association. (1994). Publication manual of the American Psychological Association (4th ed.). Washington, DC: Author.

Biskin, B.H. (in press). Comment on significance testing. Measurement and Evaluation in Counseling and Development.

Campbell, N. (1982). Editorial: Some remarks from the outgoing editor. Journal of Applied Psychology, 67, 691-700.

Carver, R. (1978). The case against statistical significance testing. Harvard Educational Review, 48, 378-399.

Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45(12), 1304-1312.

Cohen, J. (1994). The earth is round ($p < .05$). American Psychologist, 49, 997-1003.

Falk, R., & Greenbaum, C.W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. Theory & Psychology, 5(1), 75-98.

Frick, R.W. (1996). The appropriate use of null hypothesis testing. Psychological Methods, 1, 379-390.

Hagen, R.L. (1997). In praise of the null hypothesis statistical test. American Psychologist, 52, 15-24.

Harlow, L.L., Mulaik, S.A., & Steiger, J.H. (Eds.). (1997). What if there were no significance tests?. Mahwah, NJ: Erlbaum.

Harris, M.J. (1991). Significance tests are not enough: The role of effect-size estimation in theory corroboration. Theory & Psychology, 1, 375-382.

Heldref Foundation. (1997). Guidelines for contributors. Journal of Experimental Education, 65, 287-288.

BEST COPY AVAILABLE

Henkel, C.G. (1976). Tests of significance. Thousand Oaks, CA: Sage.

Hudson, W.D. (1969). The is/ought question. London: MacMillan.

Hunter, J.E. (1997). Needed: A ban on the significance test. Psychological Science, 8(1), 3-7.

Kirk, R.E. (1972). Statistical issues. Monterey, CA: Brooks/Cole.

Kirk, R.E. (1996). Practical significance: A concept whose time has come. Educational and Psychological Measurement, 56(5), 746-759.

Kupfersmid, J. (1988). Improving what is published: A model in search of an editor. American Psychologist, 43, 635-642.

Loftus, G.R. (1993). Editorial comment. Memory & Cognition, 21(1), 1-3.

Loftus, G.R. (1994, August). Why psychology will never be a real science until we change the way we analyze data. Paper presented at the annual meeting of the American Psychological Association, Los Angeles.

Mahoney, M.J. (1976). Scientist as subject: The psychological imperative. Cambridge, MA: Ballinger.

Meehl, P.E. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L.L. Harlow, S.A. Mulaik & J.H. Steiger (Eds.), What if there were no significance tests? (pp. 393-426). Mahwah, NJ: Erlbaum.

Murphy, K.R. (1997). Editorial. Journal of Applied Psychology, 82, 3-5.

Nunnally, J. (1960). The place of statistics in psychology. Educational and Psychological Measurement, 20, 641-650.

Reichardt, C.S., & Gollob, H.F. (1997). When confidence intervals should be used instead of statistical significance tests, and vice versa. In L.L. Harlow, S.A. Mulaik & J.H. Steiger (Eds.), What if there were no significance tests? (pp. 259-284). Mahwah, NJ: Erlbaum.

Rosenthal, R. (1991). Effect sizes: Pearson's correlation, its display via the BESD, and alternative indices. American Psychologist, 46, 1086-1087.

Rosnow, R.L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. American Psychologist, 44, 1276-1284.

Rozeboom, W.W. (1960). The fallacy of the null hypothesis significance test. Psychological Bulletin, 57, 416-428.

Rozeboom, W.W. (1997). Good science is abductive, not hypothetico-deductive. In L.L. Harlow, S.A. Mulaik & J.H. Steiger (Eds.), What if there were no significance tests? (pp. 335-392). Mahwah, NJ: Erlbaum.

Schmidt, F. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. Psychological Methods, 1(2), 115-129.

Schmidt, F.L., & Hunter, J.E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L.L. Harlow, S.A. Mulaik & J.H. Steiger (Eds.), What if there were no significance tests? (pp. 37-64). Mahwah, NJ: Erlbaum.

Shaver, J. (1985). Chance and nonsense. Phi Delta Kappan, 67(1), 57-60.

Shea, C. (1996). Psychologists debate accuracy of "significance

test." Chronicle of Higher Education, 42(49), A12, A16.

Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. Journal of Experimental Education, 61, 334-349.

Snyder, P.A., & Thompson, B. (in press). Use of tests of statistical significance and other analytic choices in a school psychology journal: Review of practices and suggested alternatives. School Psychology Quarterly.

Strike, K.A. (1979). An epistemology of practical research. Educational Researcher, 8(1), 10-16.

Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. Journal of Experimental Education, 61, 361-377.

Thompson, B. (1994). Guidelines for authors. Educational and Psychological Measurement, 54, 837-847.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. Educational Researcher, 25(2), 26-30.

Thompson, B. (1997a). Editorial policies regarding statistical significance tests: Further comments. Educational Researcher, 26(5), 29-32.

Thompson, B. (1997b, August). If statistical significance tests are broken/misused, what practices should supplement or replace them?. Invited paper presented at the annual meeting of the American Psychological Association, Chicago.

Thompson, B. (in press). Review of What if there were no significance tests?. Educational and Psychological Measurement, 58(2).

BEST COPY AVAILABLE

Thompson, B., & Snyder, P.A. (1997). Statistical significance testing practices in the *Journal of Experimental Education*. Journal of Experimental Education, 66, 75-83.

Thompson, B., & Snyder, P.A. (in press). Statistical significance and reliability analyses in recent *JCD* research articles. Journal of Counseling and Development.

Vacha-Haase, T., & Nilsson, J.E. (in press). Statistical significance reporting: Current trends and usages within MECD. Measurement and Evaluation in Counseling and Development.